O‘ZBEK TILI KORPUSINI YARATISH: MUAMMOLAR HALQASI VA YECHIMLAR

Javlon Jo‘rayev
j.juraev@wiut.uz
“Savodxon.uz” loyihasi asoschisi
Toshkent xalqaro Vestminster universiteti o‘qituvchisi,
“Iqtisodiy boshqaruv va rivojlanish” yo‘nalishida magistrant (MA)

Annotatsiya: O‘zbek tili uchun kompyuter lingvistikasi rivojlanishi yo‘lida to‘g‘onoq bo‘lib turgan eng katta muammo – sifatli til korpusi mavjud emasligi. Muammoning ildizi ochiq manbalardagi matn imloviy sifati pastligidir. Korpus masalasiga kirishishdan avval ana shu kamchilikni bartaraf etish zarur. Korpusni yaratish jarayoniga mutaxassislar va keng jamoatchilikni jalb qilish lozim.

Kalit so‘zlar: *til korpusi, imlo, til siyosati*

Axborot texnologiyalari shiddat bilan rivojlanib borayotgan bugungi kunda alohida bir soha taraqqiyotini an’anaviy tushunchalar doirasidagina tasavvur qilishga urinish o‘ziga xos soddadillik bo‘lib ko‘ringani bilan, uzoq muddatda, jiddiy muammolarga olib kelishi mumkin. Bugun barcha ilg‘or jamiyatlarda til fenomeni to‘liq yoki qisman dasturlashtirib bo‘lingan, hamda til va tilshunoslik eng so‘nggi texnologiyalar asosida rivojlanmoqda. To‘liq va sifatli til korpusini yaratmay turib, u jamiyatlar bunday jadal ildamlashga erisha olmasdilar [1].

Biz bugun, bir oz kechikib bo‘lsa ham, shu yo‘ldan bormoqchi ekanmiz, bunday katta vazifaga kirishishni tilga aloqador kompyuter texnologiyalari rivojlanishi uchun eng muhim xomashyo bo‘lgan *til korpusini* yaratishdan boshlashimiz kerak. Xo‘sh, bugungi kungacha o‘zbek tili korpusini yaratish uchun nimalar qilindi, bunda qanday muammolarga duch kelindi va bu muammolarga ehtimoldagi yechimlar qanday? Ushbu maqola aynan shu savollarda baholi qudrat javob berishga urinadi.

Internet tarmog‘ida izlash o‘zbek tili korpusi uchun faqat ikki manbani ko‘rsatadi: Germaniyadagi nashriyot loyihasi [2] va Chexiyadagi xususiy kompaniya mahsuloti [3]. Birinchi manba ijtimoiy loyiha sifatida olib borilgani sabab, undagi korpusni bepul yuklab olish va ishlatish mumkin. Ikkinchi manbada taklif qilingan korpusni ishlatish uchun sotib olish zarur. Ammo ikki korpus ham bir xil kamchilikka ega: unda yig‘ilgan so‘z va gaplar imloviy xatolardan holi

emas. Bu korpuslar o‘zbekcha Vikipediya, yangilik va rasmiy saytlardan yig‘ilgan bo‘lib, ana shu matn manbalarining o‘zidagi barcha xato va kamchiliklar korpusda ham aks etgan.

Kompyuter lingvistikasi mutaxassislari yaxshi biladilarki, imloviy sifati past bo‘lgan korpus uning asosida yaratiladigan dasturiy mahsulotlar sifatiga salbiy ta’sir qiladi. Masalan, neyron tarmoqlar asosida ishlaydigan sun’iy intellektni xatolari ko‘p matn bilan mashq qildirilsa, u ana shu matndagi xatolarni tilning ajralmas qismi deb qabul qilishga va shundan kelib chiqib ishlashga o‘rganadi. Oddiy misol: sun’iy intellektga berilgan korpusda “tatbiq” so‘zi 40% hollarda “tadbiq” deb yozilgan bo‘lsa, u shu so‘z o‘zbek tilida ikki xil usulda yozilar ekan deb xulosa chiqaradi va shu xulosa asosida ishlaydi. Bu juda soddalashtirilgan misol bo‘lgani bilan, uning o‘zi lingvistik dasturni “o‘qitish” uchun ishlatilgan korpus sifati uning yakuniy ishlash sifatiga qanday ta’sir qilishini tushunib yetish uchun yetarli.

Natijada, biz bugun bir-biriga tobe bo‘lgan ikki muammo orasida qolib, kompyuter lingvistikasida jiddiy siljish qila olmayapmiz. Sifatli korpusi yo‘qligi sabab, matn bilan ishlash (masalan, matndagi imlo xatolarini to‘g‘rilash) uchun yaratilgan dasturlarimiz sifati past bo‘lib qolmoqda. Tahrir dasturlari sifati pastligi, o‘z navbatida, ochiq manbalarimizdagi matnning imloviy sifatini ko‘tara olmayapti va ular asosida yaratish mumkin bo‘lgan korpus sifati pastligicha qolyapti.

Xo‘sh, bu muammolar halqasidan qanday chiqish mumkin? Avvalo, yuqori sifatli o‘zbek tili korpusini yaratish masalasida jiddiy siyosiy iroda ko‘rsatish kerak. Yuqorida keltirilgan ikki tashkilotlar uchun o‘zbek tili katta ahamiyatga ega emas – ular chet ellik mutaxassislar va ular eng avvalo o‘z ona tili yoki dunyoning katta tillariga ko‘proq e’tibor qaratadilar. Bu tabiiy hol. Tilimiz uchun to‘liq va sifatli korpus yaratish – faqat va faqat bizga kerak. Shunday ekan, bu masalaga davlat darajasida jiddiy e’tibor berish vaqti keldi. Eng muhimi, masalaga yondashuv tizimli va metodik bo‘lishi lozim.

Birinchi navbatda, o‘zbek tili imlo qoidalarini bir standartga keltirish, alifbo masalasini uzil-kesil hal qilish kerak. Alifbo masalasida siyosiy-emotsional omillardan holi bo‘lib, pragmatik yondashuvni tanlash kerak. Chunki texnologiya mafkura, milliy g‘urur yoki jamiyatdagi kayfiyat kabi subyektiv omillarga qarab rivojlanmaydi.

Alifbo va imlo qoidalari bir qolipga solingandan so‘ng, ularni amalga tatbiq qilishda siyosiy qat’iyat ko‘rsatish lozim. Nafaqat islohotdan keyin yaratiladigan matn yangi standartga mos bo‘lishini talab qilish, balki avval yaratilgan matnlarni

ham yangi standartga to‘liq o‘tkazish uchun umumiy safarbarlik e’lon qilish lozim. Shu urinishlar paytida mavjud matnlardagi kamchiliklarni biryo‘la tuzatib ketish kerak bo‘ladi.

Bunda sun’iy intellekt asosida emas – dasturiy qoidalar asosida ishlaydigan imloviy dasturlarni yaratish va matn bilan ishlaydigan barcha mutaxassislar ishiga joriy qilish kerak. Bunday dasturlarga namunalar tijorat [4] va ijtimoiy [5] loyiha ko‘rinishida ishlab chiqilgan va iqtisodiy jihatdan yangi dasturlar yaratishdan ko‘ra, shu mavjud loyihalarni qo‘llab-quvvatlash maqsadga muvofiq.

Nihoyat, sifatli til korpusini yaratish masalasiga kelsak, bunga ikki xil usulda yondashish mumkin. Birinchisi, yuqorida sanab o‘tilgan choralar natijasida yuqori sifatli matnga ega ochiq manbalar asosida xususiy sektor yuqori sifatli korpus yaratishi va uni omma bilan ulashishini rag‘batlantirish. Bu yondashuvning afzalligi shundaki, vazifani turli mutaxassislar bajarishga uringani sabab, natijada turli xil yo‘nalishdagi va jami qamrovi ancha keng bo‘lgan korpuslar to‘plamiga ega bo‘linadi. Ammo bu yondashuvda yaratib, ommaga taqdim etilayotgan korpuslar sifati tegishli darajada ekanini tekshirib turishga to‘g‘ri keladi.

Ikkinchi yondashuvda, davlat o‘z resurslarini ishga solgan holda markazlashgan chora-tadbirlar bilan yagona korpusni yaratadi va ommaga taqdim etadi. Bu yondashuvning afzalligi shundaki, unda yaratilayotgan korpus sifatini nazorat qilish va uni yagona standartga solish ancha oson bo‘ladi. Ammo bu yondashuv byudjetga qimmatroqqa tushadi.

Har ikki holatda ham yakuniy mahsulotni ommaga bepul taqdim etish kerakligi bejiz ta’kidlanmadi. Tilga aloqador dasturlar yaratishga qiziqqan professional va havaskor dasturchilarga sifatli korpus (ya’niki, dasturlashda ishlatish uchun sifatli xomashyo) bepul taqdim etilsagina, jamiyatdagi yashirin iste’dodlar o‘zini to‘liq namoyon qiladi va soha rivojlanishi keskin jadallashadi. Faqat alohida shaxslarga yaratilgan korpusdan foydalanish huquqi berilsa, korpusdan olinishi bo‘lgan naf ana shu bir guruh mutaxassislar bilim va topqirligi bilan cheklanib qoladi. Aksincha, korpusdan omma erkin foydalana olsa, chekka bir tumandagi havaskor yosh dasturchi uning asosida boshqalarning hayoliga kelmagan yangicha dasturlar yaratishi va shu bilan butun jamiyatga ko‘proq naf keltirishi mumkin.

Biroq masalaning boshqa tomoniga ham e’tibor berish lozim. Umumiy sa’y-harakatlar bilan yaratilgan til korpusidan foydalanib yaratilgan dasturiy mahsulotlar bepul bo‘lishi kerakmi? Bir qarashda, shunday bo‘lishi eng adolatli yondashuvdek ko‘rinishi mumkin. Ammo mustaqil dasturchilarga bunday talab

qo‘yilishi ulardagi yangilik yaratishga bo‘lgan intilishni kamaytiradi va jarayon samarasini tushiradi. Bunda dasturchi mehnatini yetarlicha rag‘batlantirgan holda, yakuniy dasturiy mahsulotdan butun jamiyat foydalana olishini ta’minlash tizimini ishlab chiqish va joriy qilish zarur. Bunday tizim qanday bo‘lishi kerakligi ushbu maqola mavzusidan tashqaridagi masala.

Xulosa qilib shuni aytish mumkinki, o‘zbek tili uchun kompyuter lingvistikasi soha sifatida taraqqiy etishi yo‘lida turgan to‘siqlar aniq, ulardan ko‘pchilik boxabar va ularni ko‘p mutaxassislar ta’kidlab o‘tgan (ushbu maqola bu muammoni ko‘targan birinchi yoki yagona manba emas). Ushbu maqola ana shu muammolarni yana bir bor soha mutaxassislari muhokamasiga olib chiqish, til siyosatiga javobgar tashkilotlar e’tiborini ularga qaratishga urindi. Unda mavjud muammolarning eng kattasi – o‘zbek tilining sifatli korpusi haligacha mavjud emasligiga tegishli yechimlar taklif etildi. Bu yerda va bundan avval taklif qilingan yechimlar hayotga qanchalik samarali tatbiq qilishini esa vaqt ko‘rsatadi.

Foydalanilgan adabiyotlar:

1. G‘ulomova, (2020): https://uzjournals.edu.uz/buxdu/vol4/iss4/9/
2. Springer, Berlin, Heidelberg: https://cls.corpora.uni-leipzig.de/en?corpusLanguage=uzb#tblselect
3. Lexical Computing CZ s.r.o.: https://www.sketchengine.eu/corpora-and-languages/uzbek-text-corpora/
4. Savodxon.uz loyihasi: https://savodxon.uz/
5. Matn.uz loyihasi: https://matn.uz/